CARNAC-LR : Clustering coefficient-based Acquisition of RNA Communities in Long Reads

Camille Marchet, Lolita Lecompte, Corinne da Silva, Corinne Cruaud, Jean-Marc Aury, Jacques Nicolas, Pierre Peterlongo

CARNAC-LR : Clustering coefficient-based Acquisition of RNA Communities in Long Reads

Camille Marchet*†[1], Lolita Lecompte[1], Corinne Da Silva[2], Corinne Cruaud[2], Jean-Marc Aury[2], Jacques Nicolas[1], and Pierre Peterlongo[1]

[1]Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA) – Universite de Rennes 1, Institut National de Recherche en Informatique et en Automatique – Avenue du général LeclercCampus de Beaulieu 35042 RENNES CEDEX, France

[2]Genoscope-Centre national de séquençage (GENOSCOPE) – CEA – Genoscope-Centre National de Séquençage 2, rue Gaston Crémieux CP5706 91057 EVRY Cedex, France

Résumé

Motivation

Lately, long read sequencing technologies, referred to as Third Generation Sequencing, (TGS, Pacific Bioscience [1] and Nanopore [2]) have brought the opportunity to sequence full-length RNA molecules. In doing so they relax the constraint of transcript reconstruction prior to study complete RNA transcripts. By avoiding limitations of previous technologies, [3, 4] and giving access to the trancripts structure, they might contribute to complement and improve transcriptomes studies. This is particularly crucial for non model species where assembly was required. Many biological questions (finding gene signatures for a trait, finding expressed variants...)[5, 6] are classically addressed using transcriptome sequencing. However, this gain in length is at the cost of a computationally challenging error rate (up to 15%) that disqualifies previous short-reads methods. In this work we propose to support the analysis of RNA long read sequencing with a clustering method that works at the gene level. It enables to group transcripts that emerged from a same gene. From the clusters, the expression of each gene is obtained and related transcripts are identified, even when no reference is available.

Problem statement

Within a long reads data set, our goal is to identify for each gene the whole set of reads that come from its expression without the help of a reference genome or transcriptome. This problem can be computationally formalized as a community detection problem, where a community (also referred to as a cluster) is the population of reads coming from a same gene. Communities are densely connected groups of nodes, although there exists no rigorous shared definition. Our application problem is non trivial and specific for three reasons:

1-in eukaryotes, it is common that alternative spliced and transcriptional variants (called isoforms) which differ in exon content occur for a given gene. In this case we want alternative transcripts to be grouped in a same cluster;

*Intervenant

†Auteur correspondant: camille.marchet@irisa.fr

2- long reads currently suffer from high error rates and computationally challenging error profiles with a majority of indels errors;

3- all genes are not expressed at the same level in the cell, which leads to an heterogeneous coverage in reads of the different genes, then to communities of different sizes including small ones. This can be a hurdle for community detection.

Previous works

The problem in itself is not new, it dates back before the advent of NGS, with Sanger sequencing and the necessity to cluster ESTs. However these methods were tailored to work with lower scalabilty challenges due to the scarcity of data, and a far less important error rate than with current long reads. The concept of community detection is a natural way of depicting our problem. Due to the ambiguity of the community definition, a plethora of methods have been proposed for their detection. Moreover this problem has appeared in many disciplines, taking many slightly different forms according to the application domain. The first approach that brought an important contribution is an algorithm based on *modularity*. Other methods were then proposed as improvements, in particular methods relaxing the definition of communities as objects that can overlap, such as the Clique Percolation Method (CPM) [7].

Contribution

Roughly speaking, resolution strategies can be classified into two trends according to applications and the community of affiliation: a *graph clustering* strategy based on the search for minimal cuts in these graphs and a *community finding* strategy based on the search for dense subgraphs. Our own approach aims to combine the best of both worlds.
The first approach generally searches for a partition into a fixed number of clusters by deleting a minimum number of links that are supposed to be incorrect in the graph. The second approach frequently uses a *modularity* criterion to measure the link density and decide whether overlapping clusters exist, without a priori regarding the number of clusters. Given that it is difficult to decide on the right number of clusters and to form them solely on the basis of minimizing potentially erroneous links, the main findings and recent developments are based on the community finding strategy and we will focus our review on this approach.
Our algorithm is based on the concept of *clustering coefficient* and we formalize our problem as finding communities such that a community is a connected component in the graph of similarity having a *clustering coefficient* above a fixed cutoff, and such that communities are disjoint sets. An optimal clustering in k communities is a minimal k-cut of the graph, that is, a set of k disjoint subsets of reads such that the set of edges between two different subsets has minimal size. We then implement heuristics that approximate a result of this problem. They are implemented in a tool dubbed CARNAC-LR (Clustering coefficient-based Acquisition of RNA Communities in Long Reads), integrated into a pipeline. The input is a set of long reads and the output is a file with reads indexes grouped in one line per cluster. Our approach is compared to state of the art algorithms to detect communities. We then show its relevance on a real data set issued from mouse brain transcriptome.

Results
We use a real mouse dataset sequenced using a MinION platform at the Genoscope to compare our solution to other algorithms used in the context of biological clustering and demonstrate its is better-suited for transcriptomics long reads.
We build "ground truth" clusters using mapping routine that are compared to *de novo* clustering results.
We use them to benchmark classic community detection algorithms and state of the art sequence clustering tools such as CD-HIT [8] and show we perform better on ONT reads from mouse.
When a reference is available thus mapping possible, we show that it stands as an alternative

method that predicts complementary clusters.

References

Manuel L Gonzalez-Garay. Introduction to isoform sequencing using pacific biosciences technology
(iso-seq). In Transcriptomics and Gene Regulation, pages 141–160. Springer, 2016.

Jason L Weirather, Mariateresa de Cesare, Yunhao Wang, Paolo Piazza, Vittorio Sebastiano, Xiu-Jie Wang, David Buck, and Kin Fai Au. Comprehensive comparison of pacific biosciences and
oxford nanopore technologies and their applications to transcriptome analysis. F1000Research, 6, 2017.

Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido
Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, et al. Trinity: reconstructing
a full-length transcriptome without a genome from rna-seq data. Nature biotechnology, 29(7):644, 2011.

Tamara Steijger, Josep F Abril, Par G Engstrom, Felix Kokocinski, Tim J Hubbard, Roderic Guigo, Jennifer Harrow, Paul Bertone, RGASP Consortium, et al. Assessment of transcript reconstruction methods for rna-seq. Nature methods, 10(12):1177–1184, 2013.

Robert Ekblom and Juan Galindo. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity, 107(1):1, 2011.

Chien-Yueh Lee, Yu-Chiao Chiu, Liang-Bo Wang, Yu-Lun Kuo, Eric Y Chuang, Liang-Chuan Lai, and Mong-Hsun Tsai. Common applications of next-generation sequencing technologies in
genomic research. Translational Cancer Research, 2(1):33–45, 2013.

Gergely Palla, Albert-Laszlo Barabasi, and Tamas Vicsek. Quantifying social group evolution.
arXiv preprint arXiv:0704.0744, 2007.

Li, Weizhong, and Adam Godzik. "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." *Bioinformatics* 22.13 (2006): 1658-1659.

Mots-Clés: transcriptomics, oxford nanopore technologies, long reads, simulation, mRNA